Big Data Intelligence HW2

2024-10-29

Problem: Personalized Recommendation

- Recommendation is a typical application of big data.
- Using the known browsing history of users, predict the interests of users and recommend new information.
- Given user behavior matrix x, x is a matrix of $m \times n$, where m is the number of users and n is the number of movies. Given some of the values in x, how can we guess the unknown values?
- Dataset： Netflix
  - Film rental provider, one of "FAANG".
  - In 2006, the Million Dollar Data Challenge was set up with the goal of improving the recommendation accuracy by 10%.

Data set: a Subset of Netflix Recommendation Contest.

- Users rate movies, with ratings ranging from 1 to 5, with a total of 10,000 users and 10,000 movies. 80% of the behavior data is the training set, and the remaining 20% is the test set.
  - `users.txt`: 10000 lines, each line is an integer, which indicates a user id for a user. This file corresponds to all users.
  - `movie_titles.txt`: The movie information. Each line corresponds to movie id, release year, and movie name.
  - `netflix_train.txt`: The training set. It contains 6.89 million user ratings, where each line is one rating, includes user id, movie id, rating and rating date. The user id and movie id is consistent with the one in `users.txt` and `movie_titles.txt`.
  - `netflix_test.txt`: The testing set, including 1.72 million user scores, with the same format as the training set.

Experiment 1：Data Preprocessing

- Organize the input file into a matrix X with the dimension of [user_num, movie_num], where:$\mathrm{X}[i,j]$ is the rating of the movie j given by the user i. Output two matrices:$\mathrm{X}_{train}$and X_{test}, which correspond to the training and testing set, respectively.
- If it is difficult to deal with the matrix of 10000×10000, you are allowed to sample a subset of the matrix composed of some users and some, but it must not be less than 2000*2000. It is also necessary to explain the sampling rules in the report. Note that sampling a subset will lead to a small score discount for your homework.

Experiment 2: Collaborative Filtering

- Implement user-based collaborative filtering algorithm: predict whether user i likes movie j, based on the ratings of users similar to user i and see how they like item j. Also, the similarity between user k and user i could be used to reweight the score of user k on movie j.

$$\text{score(i, j)} = \frac{\sum_k sim\big(X(i), X(k)\big) \cdot score(k, j)}{\sum_k sim(X(i), X(k))}$$

- The similarity between user i and user k in scoring movies can be expressed by $cosine$ similarity of two vectors.
- Evaluation Metric：RMSE(Rooted Mean Square Error)

$$\text{RMSE} = \sqrt{\frac{1}{|Test|}\left(\sum_{\langle i,j\rangle \in Test} (X_{ij} - \tilde{X}_{ij})^2\right)}$$

Experiment 3: Matrix Decomposition

- Implement the matrix decomposition algorithm based on gradient descent : the behavior matrix X is decomposed into the product of U and V, so that $U \cdot V$ approaches X at the known value: $\mathrm{X_{m*n}} \approx \mathrm{U_{m*k}} {V_{n*k}}^{T}$
- The hidden dimension k is the parameter of the algorithm, U and V can be regarded as the characteristic expressions of users and movies in hidden space, and its product matrix can predict the unknown part of X.
- Optimal objective $J = \frac{1}{2}\|A \odot (X - UV^T)\|_F^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2$
- Gradient descent optimization
- Learning rate and convergence

Initialize U and V
Loop until convergence
    Calculate $\frac{\partial J}{\partial U}, \frac{\partial J}{\partial V}$
    Update $U = U - \alpha\frac{\partial J}{\partial U}, V = V - \alpha\frac{\partial J}{\partial V}$
End loop

Experiment 4 (Optional)

- Features like user rating date and movie name in the dataset have not been used up to now.
- How to use this extra information to improve the recommendation algorithm, so as to further improve the recommendation performance? How important is this extra information to the recommendation?
- For students who are interested in the optional content, please write down the proposed method, corresponding experimental results and relevant analysis into the experimental report, and attach the corresponding codes, which will bring bonus points.

Some Notes

- Homework Duration: 4 weeks. Submission requirements: see pdf for details.
- In this experiment, X matrix is very sparse, so we recommend that you use sparse matrix operation, which will be faster. The reference formula provided in the collaborative filtering section is not in matrix form, so please deduce it by yourself.
- RMSE calculation: based on the problem background (the rating is 1-5 points) and the definition of RMSE, how large RMSE is reasonable? What is the RMSE with random guess?
  - If the result of your algorithm is not as good as random guess, there must be an error …
- Remind again: Start as soon as possible, and no Plagiarism.